Compositional analysis approach in the measurement of social-spatial segregation trends. A case study of Guadalajara, Jalisco, Mexico.

M.A. Cruz[1], M.I. Ortego[1], and E. Roca[1]

[1]Department of Civil and Environmental Engineering.
Universitat Politècnica de Catalunya BarcelonaTech, Spain;
marco.antonio.cruz@upc.edu

Abstract

The place in which we live affects our outlook on life; it is through the place of residence and its surroundings where our relationships, thoughts and opportunities are born and shaped. Different authors have highlighted the internal existing social differences in cities as a consequence of the neoliberal system in which we live. The mercantile logic that affects urban spaces incentives the dichotomy winners-losers in the current urban landscape and leads to the differentiation and unequal distribution of certain social groups within the urban space. This clear differentiation in distribution of social groups in the urban space has been called socio-spatial segregation. This concept arises from the urban sociology, the first studies were focused on the differentiation of ethnicity and income level to identify the most vulnerable groups and of mitigate their current situation through different policies.

A more significant number of variables belonging to different dimensions (social, economic, political and environmental) have been incorporated into the study of this phenomenon, traditionally addressed by different disciplines such as sociology, geography and anthropology. Nonetheless, few studies have addressed it from a multivariate analysis approach. Moreover, the few existing studies with a multivariate statistical analysis ignored or did not know the compositional nature of their data.

The objective of the present study is to introduce the compositional data analysis in urban studies to better understand socio-spatial segregation in the different urban contexts. Specifically, the analysis of social-spatial segregation considering the compositional nature of the data in the city of Guadalajara, Mexico, is carried out. Socio-economic variables from census data of approximately 13,520 urban blocks grouped in 395 colonias and seven urban districts are used to carry out this study through the most straightforward compositions of two components. Additionally, principal component analysis and cluster analysis are performed to identify the socio-economic distribution within the territory. The analysis is complemented with the use of geographic information systems (GIS) at different urban scales.

Based on Aitchison log ratio approach, the results are consistent with the segregation processes that date back to the foundation of the city. Through cluster analysis and principal component analysis, an evident polarization between the Minerva district and the rest of the areas is shown.

Key words: Compositional analysis, segregation, Guadalajara.

1 Introduction

In its broader conception, the existence of differentiation or unequal distribution of certain social groups within the urban space is known as urban segregation (Brun 1994). The term segregation emerged from urban sociology, and its study is traced back to the first half of the 20th century by the Chicago School of Sociology (Dawkins, Reibel and Wong 2007). From its origins, the social division of space was strictly linked to the concepts of income and race. However, nowadays when talking about segregation it is necessary to broaden the spectrum of elements that generate it and treat it as a dynamic process caused by different factors with different intensities that can accelerate or sustain the phenomena of segregation (Donat 2018); (Massey and Denton 1988); (Peach 1996); (Subirats 2004).

Although the study of segregation began around 1920, the first measurement is attributed to Jahn in 1947 with the index of dissimilarity (Jahn, Schmid and Schrag 1947). In his work, the author highlighted the applicability of his index, which can be applied to any population or class. However, according to Cowgill and Cowgill (1951), this index had two major flaws. The first of them, its difficulty to calculate it and the second, a lack of precision in the measurement differences in concentration and dispersion.

The interest of different authors for this phenomenon, coupled with technological advances, led to the appearance of a large number of indexes. For this reason, Massey and Denton (1988) emphasized on the existing state of theoretical-methodological disorder and a minor consensus in the use of indicators by the researchers of that time. In their study, the twenty most relevant existing segregation indexes in the literature were selected and classified in five dimensions: uniformity, exposure, concentration, centralization and clustering.

Concerning the first dimension, Massey and Denton (1988) proposed the dissimilarity index of Duncan and Duncan (1955), which measures the degree to which the groups are distributed differently in the urban space. The Lieberson index (1981) with respect to the exposure dimension was selected to represent the degree to which the members of different groups share common areas. The index of relative concentration was proposed to measure the degree of agglomeration of a group in the urban space within the dimension of concentration. On the other hand, the absolute centralization index was chosen to evaluate the degree to which the groups are located in the center of the cities. Finally, the spatial proximity index was selected to measure the proximity between groups in the urban space within the dimension of clustering.

The lack of the element of spatiality in the measurement of segregation represented the greatest criticism for the former indexes (Romero Mares and Hernández Lozano 2015); (Johnston, Poulsen and Forrest 2015); (Wong 2015); (Lloyd, Catney and G.Shuttleworth 2015). Likewise, the proposed indexes in their different dimensions simplified the measurement of segregation in considering an average value applied to race and religion as a total indicator of segregation of a territory without showing the various nuances that may exist within it (Johnston, Poulsen and Forrest 2015).

The ambiguity of the concept of segregation (Duncan and Duncan 1955); (Romero Mares and Hernández Lozano 2015); (Donat 2018), together with the incompatibility in the census information (temporality, statistical geographical areas and indicators), has made impossible the evaluation and comparison between cities and population groups in the different national contexts (Kertzer and Arel 2002); (Mateos 2015).

Over time, different methods for measuring segregation have been proposed. These methods have evolved based on the disaggregation of census information, statistics, the dimensions of study and the available technologies for the treatment of data (Lloyd, Shuttleworth and Wong 2015); (Romero Mares and Hernández Lozano 2015). Nonetheless, few studies have addressed it from a statistical multivariate analysis approach (Schteingart 2015). Moreover, the few existing studies with a multivariate statistical analysis ignored or did not know the compositional nature of their data such as the ones elaborated by CONAPO (2010), Romero and Hernández (2015), Shteingart (2015) and Jiménez and Donat (2018) among others.

For such a motive, the objective of this study is to introduce compositional data analysis in urban studies to better understand socio-spatial segregation in the different urban contexts. Specifically, this approach is applied to Guadalajara, Mexico. For this purpose, census information of approximately 13,520 urban blocks grouped into 395 colonias and seven urban districts of the city is analyzed as two-part compositions.

2 Case study of Guadalajara

Recent urban policies characterized by the hegemony of certain economic and political forces have made of Guadalajara a place where the interests of class, the logic of wealth and accumulation have overlapped rational urban planning. As mentioned by Kempen (2002) , cities are not naturally divided, its division is the product of an intentional and active act of those that have the power to do it. As regards to Guadalajara the division of space dates back from its foundation in 1542. The Spanish Crown, through its urban ordinances and know-how, established a defined geometric scheme within the urban fabric and a clear social hierarchy in the city (López Moreno 2001).

Among the criteria of the know-how by the Spanish Crown, in order to guarantee access to water, cities should be located in the proximity of a river (Vázquez 1989). In the latter sense, the city of Guadalajara settled on to one side of the San Juan de Dios River. A natural border that internally divided the city (Aceves, Torre and Safa 2004). The local bourgeoisie, the rich and renowned people, would concentrate on the west side of the river (except for the 200 indigenous allies who were responsible for the defense of the city). On the other hand, the indigenous population with no nobility titles was located east of the river, installing a clear division of the urban space, a Spanish city versus the city of Indians (Vázquez, 1989).

Even though in 1896 the San Juan de Dios River was forced underground and the Independencia Causeway was built over it, the urban planned growth continued on the west side of the river, such as the Colonias in the period 1894-1924. Product of foreign

capital and in its beginnings inhabited by foreigners and wealthy families, Colonias were homogeneous subdivisions that responded to commercial interests that sought the increase of the value of the land and that accentuated the east-west and poor-rich dichotomy of the city (Alvizo 2013); (Barajas and Muñoz 2006); (Cabrales and Canosa 2001); (Doñán 2013); (López Moreno 2001); (Rivera 2012); (Vázquez 1989).

In addition, the incorporation of neoliberal policies and the modification of structural reforms in Mexico in the period from 1982 to 1988, provoked the abandonment of the State in its role of urban planner what has meant a more significant social division and differentiation of the space in the Mexican cities (Marrufo and Bass Zavala 2015); (Moreno Pérez 2015); (Rivera 2013).

3 Methodology

To carry out the present study, a descriptive multivariate statistical analysis is performed, the compositional nature of the data and their properties are considered (Aitchison 1986); (Pawlowsky-Glahn and Egozcue 2006). This approach implies that the data used is in a restricted space called the Simplex; likewise , the data used is part of a whole, is positive, is in a range between zero and a positive number, and is subjected to the sum of a constant (Pawlowsky-Glahn and Egozcue 2006). This fact conditions the relationship that variables have to one another. Data does not vary independently as they would if they were not subject to the sum of the constant and that can be seen in the variance-covariance structure. The constant sum constraint forces at least one covariance to be negative and at least con correlation between elements will be negative. The latter means that coefficients between elements range between -1 and +1, which leads to the existence of spurious correlations.

To overcome the consequences of working with compositional data and to validate the statistical analysis, the log-ratio approach proposed by Aitchison (1986) is applied. This approach allows us to work with log ratios of compositions as if they were real random variables, and therefore the multivariate classic statistics tools can be applied. Moreover, the analysis is based on the relative information between the components and not on their absolute values (Aitchison 1986).

The indicators used in this study are simple two-part compositions, which are grouped and analyzed in different dimensions (i.e., social and economic), see Equation (1). As a result of having the sum-constraint the data is transformed with the log-ratio approach. Hence, the second component analyzed will be 100 minus the sum of the first component, see Equation (2).

$$X = C[x_1, x_2] \in S^2 \quad (1) \qquad log(X) = \frac{x_1}{1-x_1} \quad (2)$$

Where:

X = Compositional vector of two parts.

C = Stands for closure. The vector has been rescaled making the components add to 100.

x_1 , x_2 = Parts of our compositional vector.

S^2 = Subset of D=2 dimensional real space in the simplex.

Important care should be addressed in the values used in the numerator and the denominator of Equation (2). The above has a direct influence on the results obtained. Therefore, in this study, the aspects of the compositions that were considered as positive aspects were placed in the numerator and negative elements in the denominator (e.g., log ratio of educated population and uneducated population).

Once the information is presented in log ratios, a descriptive multivariate statistical analysis is carried out in R. First, a Principal Component Analysis (PCA) is done for a dimensionality reduction. This analysis explains the variability of our data through a set of new dimensions. PCA illustrates the behavior of the observations, the relationship and the direction of the variables. Moreover, it helps to identify clusters of observations in the data. Second, a hierarchical cluster analysis using Wards method and Euclidean distances is performed. Cluster analysis allows to group observations with similar behavior and allows to know the characteristics of the observations grouped in the different groups.

Finally, Geographic Information Systems (GIS) were used to incorporate the spatial element into the study. These systems allow the visualization of the distribution and the social division of the space based on the different indicators and dimensions analyzed.

4 Data

The present study includes the analysis of census information of approximately 13 520 urban blocks, which are grouped in 395 colonias of the city and which in turn are grouped into seven large urban districts. Therefore, this study involves the use of different sources of information for its realization and which are described below.

4.1 Geospatial vector information

The vector information corresponding to the geographic information systems in its shapefile format is obtained from two sources. The data from the 13,520 urban blocks and the territorial limits of the city of Guadalajara is obtained from the National Institute of Statistics and Geography (2010). On the other hand, the territorial delimitation of the 395 colonias and the seven urban districts is obtained from GeoGDL (2019).

4.2 Census information

The census information in Mexico is carried out every ten years. For this reason, the most recent census information at the urban block level corresponds to the year 2010. Information corresponding to the indicators used in this study is obtained from the 2010

Population and Housing Census (INEGI 2010). In this information, unique identification codes are presented for the different urban blocks. Therefore, the data is linked with Geographic Information Systems, which facilitates the processing of data at different scales such as colonias and districts.

4.3 Dimension and variables

Since INEGI did not include the income variable in the census of 2010, indicators of material goods and services in substitution of the income received are used as an approximation of the economic dimension, see Table (1).

Table 1 Variables measuring the economic dimension. Elaborated based on INEGI (2010)

Indicators	**Definition**
RelVPH_TV	Log ratio of private households owning a television and private households without a television.
RelVPH_RE	Log ratio of private households owning a refrigerator and private households without a refrigerator.
RelVPH_LA	Log ratio of private households owning a washing machine and private households without a washing machine.
RelVPH_AU	Log ratio of private households owning an automobile and private households without an automobile.
RelVPH_PC	Log ratio of private households owning a computer and private households without a computer.
RelVPH_TE	Log ratio of private households owning a telephone and private households without a telephone.
RelVPH_CE	Log ratio of private households owning a cellphone and private households without a cellphone.
RelVPH_IN	Log ratio of private households with internet service and private households without internet service.

5 Results

It should be noted that the census information regarding to the 13,520 urban blocks has been scaled up and grouped in the different 395 colonias. Likewise, the colonias are differentiated by utilizing the seven districts to which they belong. The representation of both scales (colonias and districts) allow a better understanding of their behavior based on the different set of dimensions.

5.1 Principal Component Analysis

A high correlation among independent variables creates problems such as multicollinearity. Accordingly, a reduction of dimensions is performed through a Principal Component Analysis (PCA). The PCA preserves as much as possible of the original structure of the data. Furthermore, this analysis generates and defines a new set of independent dimensions by taking the data set and looking for directions explaining the greatest amount of variability, see Table (2).

Table 2 Importance of components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	2.17	1.61	0.61	0.35	0.27	0.22	0.13	0.08
Proportion of variance	59.10	32.74	4.66	1.613	0.97	0.61	0.22	0.09
Cumulative proportion	59.10	91.84	96.50	98.11	99.08	99.69	99.91	100.00

Since PC1 and PC2 captures a cumulative proportion of the 91.84 % of the variability of the information, a biplot with these two components is performed. The biplot allows to identify the behavior of the colonias around the different indicators, it helps to explain the relationship between them, and at the same time, it identifies potential clusters. See Figure (1).

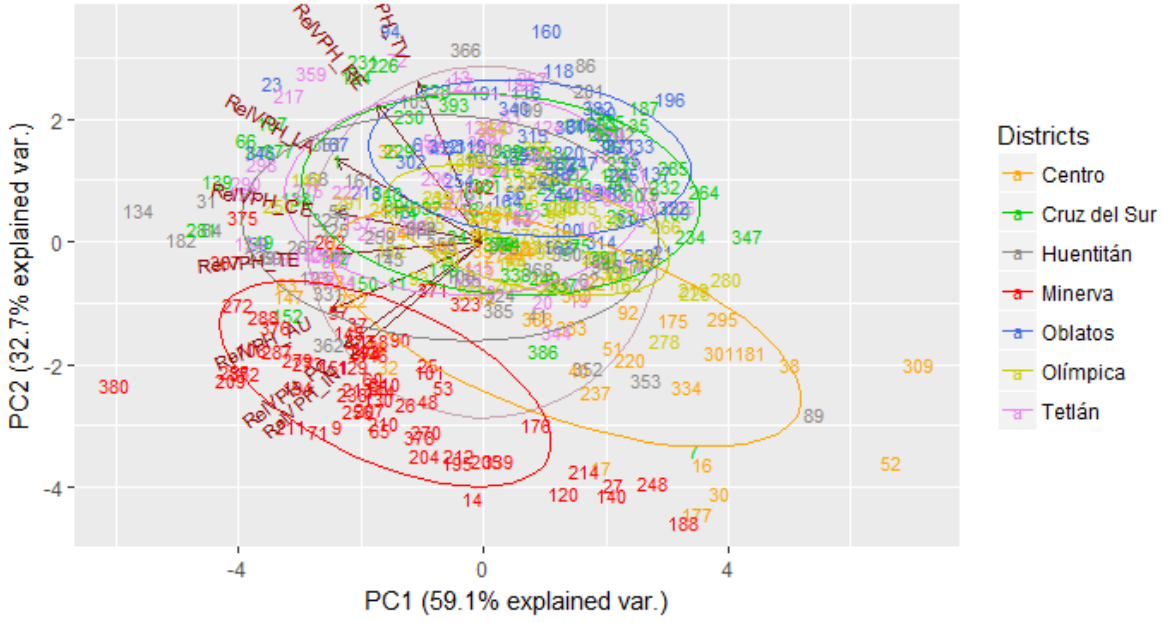

Figure 1 Principal Component biplot explaining 91.83 % of the variability of observations by colonias and urban districts.

From the biplot it is observed that the colonias belonging to the Minerva district are characterized by having a similar behavior between them and opposite from the rest. Particularly, it is observed that in the Minerva district and its colonias, the households